Evolving feature extraction algorithms for hyperspectral and fused imagery

Steven P. Brumby, Paul A. Pope, Amy E. Galbraith, and John J. Szymanski
Space and Remote Sensing Sciences,
Los Alamos National Laboratory, Mail Stop D436,
Los Alamos, New Mexico 87545, USA.
brumby@lanl.gov

Abstract *- Hyperspectral imagery with moderate spatial resolution (~30m) presents an interesting challenge to feature extraction algorithm developers, as both spatial and spectral signatures may be required to identify the feature of interest. We describe a genetic programming software system, called GENIE, which augments the human scientist/analyst by evolving customized spatio-spectral feature extraction pipelines from training data provided via an intuitive, point-and-click interface. We describe recent work exploring geospatial feature extraction from hyperspectral imagery, and from a multi-instrument fused dataset. For hyperspectral imagery, we demonstrate our system on NASA Earth Observer 1 (EO-1) Hyperion imagery, applied to agricultural crop detection. We present an evolved pipeline, and discuss its operation. We also discuss work with multi-spectral imagery (DOE/NNSA Multispectral Thermal Imager) fused with USGS digital elevation model (DEM) data, with the application of detecting mixed conifer forest.*

Keywords: Genetic Programming, Image Processing, Hyperspectral Imagery, Multispectral Imagery, Digital Elevation Model, Data Fusion, Remote Sensing.

1 GENIE feature extraction technique

Los Alamos National Laboratory's GENIE software [1-4] is a machine learning software system using techniques from the fields of genetic algorithms (GA) [5-7] and genetic programming (GP) [8] to construct feature extraction algorithms for remotely sensed imagery. Both the structure of the feature extraction algorithm, and the parameters of the individual image processing steps, are learned by the system. GENIE has been described at length elsewhere [1-4], so we will only present a brief description of the system here.

GENIE follows the paradigm of genetic programming: a population of candidate image-processing algorithms is randomly generated from a collection of low-level image processing operators, including texture measures, spectral band-math operations (e.g. ratios of bands), and various morphological filters. The fitness of each individual is assessed from its performance on training data provided by the human user via a graphical interface. Our fitness metric in based on measuring the total error rate (false positives and false negatives) on the feature extraction task. After a fitness value has been assigned to each candidate algorithm in the population, the most fit members of the population reproduce with modification via the evolutionary operators of mutation and crossover. This process of fitness evaluation and reproduction with modification is iterated until the population converges, or some desired level of classification performance is attained, or some user-specified limit on computational effort is reached (e.g. number of candidate algorithms evaluated). The final result is a grey-scale enhancement of the feature of interest, which is then converted into a final boolean classification using a threshold. This final threshold may be adjusted by the human user to take into account the desired emphasis of the value of detection rate (true positives) over false alarms and missed detections.

The algorithms evolved by GENIE combine spatial and spectral processing, and the system was in fact designed to enable exploration of spatio-spectral image processing. This system has been shown to be effective in detecting complex spatio-spectral terrain features in multispectral imagery, such as golf courses in MODIS Airborne Simulator imagery [9], and in a range of real world problems, including delineating and classifying wildfire burn scars [10] and vegetation land-cover classes [11] using a number of multispectral imagery datasets, and detecting craters on Mars [12] using a high-resolution panchromatic dataset (Mars Global Surveyor/Mars Orbital Camera). We now describe work exploring the use of this system with two challenging types of remotely sensed data: hyperspectral imagery and multi-instrument fused imagery.

2 Detecting Crops with Hyperion HSI

GENIE was originally designed to evolve feature extraction algorithms for multispectral imagery, so the extention to hyperspectral imagery was viewed as a test of the scalability of the technique (from 10's to 100's of spectral bands). Hyperspectral imagery is often analysed using purely spectral techniques such as the spectral angle

Figure 1. Visible (left) and Color-Infrared (right) views of the training region extracted from the Hyperion sample scene. Rice, soy, and corn fields are present, as well as unplanted ploughed fields, natural vegetation, and roads and buildings. At 30m spatial resolution, textural differences between crops are noticeable.

mapper (SAM) algorithm, or by the design of matched filters that use the whole spectrum of each pixel (for a general review of hyperspectral image processing, see, e.g., the textbook treatment in [13]). GENIE's set of primitive image processing operators from which it builds its candidate algorithms are designed to work on only one or two spectral bands of data, and so this experiment tests the ability of the GENIE system to exploit the inherent redundancy of HSI and identify a small number of relevant bands out of the full hyperspectral data cube.

Our hyperspectral imagery data source is a sample scene released by the Hyperion instrument team. Hyperion (see, e.g., [14] and references therein) is an experimental, moderate-resolution (~30m/pixel), 220 band visible (~0.4 μm) to short wave infrared (~2.5 μm) hyperspectral imager flown on the NASA New Millenium Program's Earth Observing 1 (EO-1) spacecraft. The scene we used covers part of the Coleambally Irrigation Area, an intensively farmed agricultural region located in the state of New South Wales, Australia (image collected 6 March, 2001). This region produces a number of commercial crops, including rice, corn and soy beans. Individual fields are large enough that even at ~30m/pixel spatial resolution, textural cues to the nature of planted crops are obvious (e.g., terracing of rice paddies).

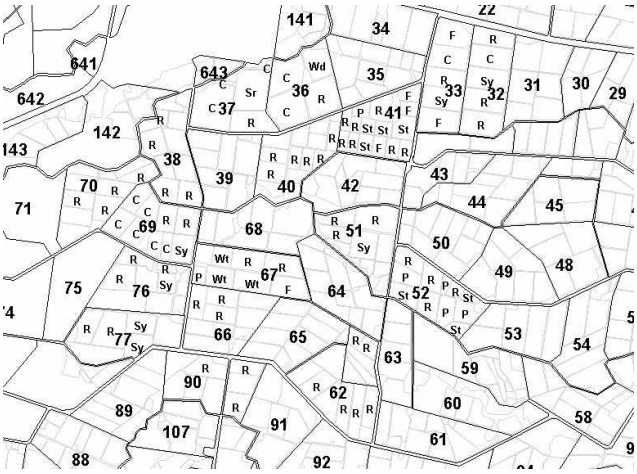

Figure 2. Ground truth for the system of fields shown in Fig. 1. Labels of interest are R: Rice, C: Corn, Sy: Soy. Proceedings of the Hyperion Data User Workshop, 2001.

Figure 1 shows our 256x256 pixel training region, which is a small part of the full (417x752 pixel) sample scene. This region was chosen because of the availability of ground truth, in the form of a map of planted fields, shown in Figure 2. For each of three crops, rice, soy, and corn, a small amount of training data was marked up, and the GENIE system was used to evolve a feature extraction algorithm for that crop (training data and result for each crop is shown in Fig. 3). Table 1 presents our detection and false alarm rates for the image on the training data.

Considering performance outside of the training data, on comparing the results in Figure 3 with the ground truth in Figure 2 we see that the algorithms evolved by GENIE have marked out physically reasonable crop regions for each of the crops of interest. These agree very well with the ground truth map of crop planting.

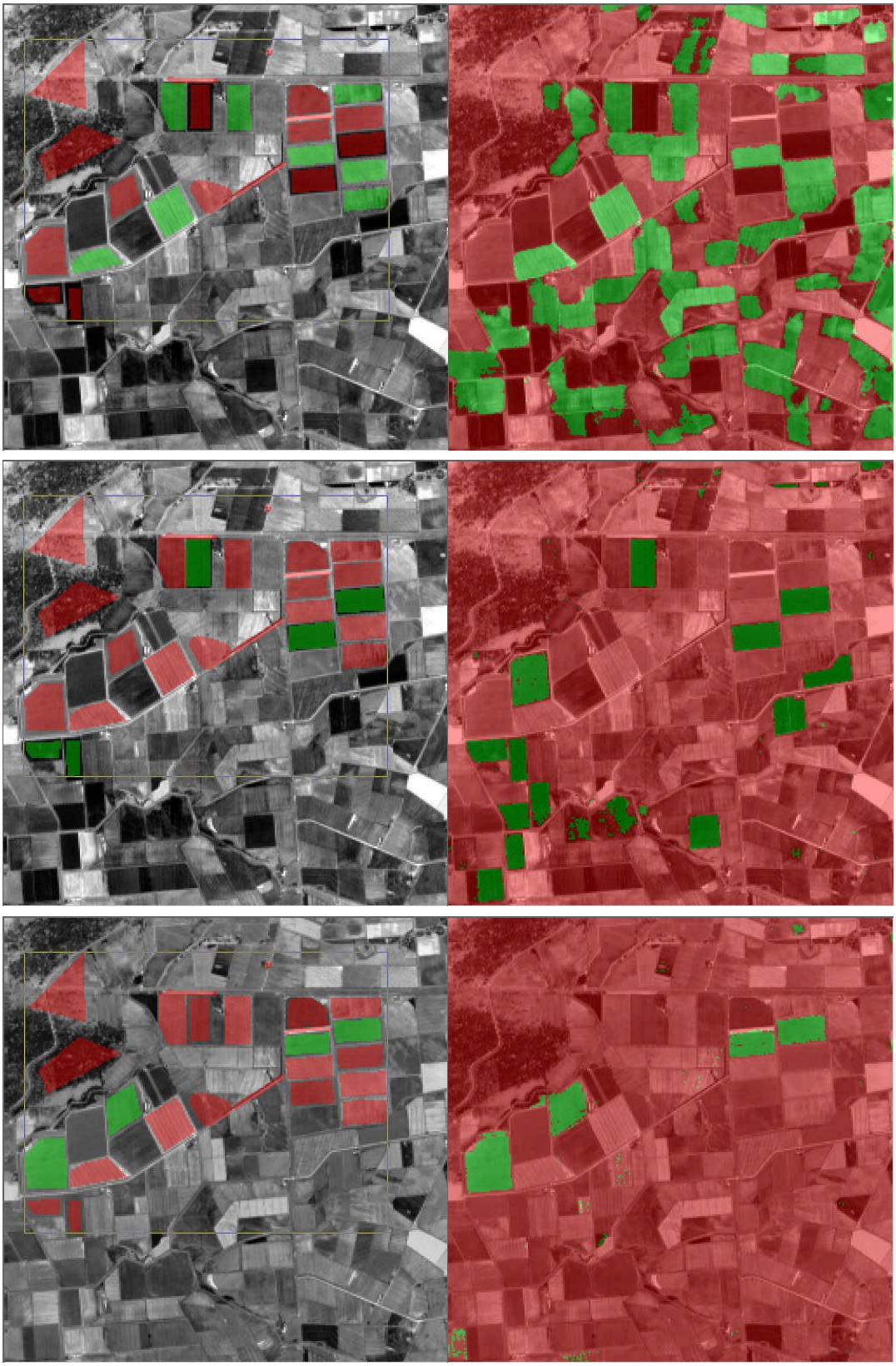

Figure 3. GENIE training data (left column) and results (right column) for rice, soy, and corn crops (from top to bottom). Green pixels mark the feature of interest, and red pixels mark background pixels. These results compare well to the ground truth presented in Figure 2.

Table 1. GENIE result for crop detection on the training data

Feature	Detection Rate [%]	False Alarm Rate [%]
Rice	95.5	0.0
Soy	99.9	0.0
Corn	98.3	0.4

In the case of Corn, fields in the south-west quadrant of field region #36 and north-west corner of field region #37 (see Fig.2) are not detected, but on inspection of the raw imagery (Fig.1) it is not clear that these fields have been correctly labeled in the "ground truth" (Fig.2). We intend to obtain clarification on this point from the Coleambally Irrigation Area managers.

The algorithms found by GENIE used both spatial and spectral processing. We will describe the soy algorithm in a little detail, as this is representative of our other results. Detail on all these algorithms, and comparison to other standard algorithms, will appear in a future publication.

The soy algorithm, which appears in our genetic programming representation as the text string (see [1,4])

[SPIKE rD175 wS1 0.34 0.85][MSAVI rD17 rD115 wS2] [R5R5 rD217 wS3]

constructs three spatio-spectral signature bands:

- An amplitude band-pass filter (SPIKE) is applied to spectral band 175 (1.901 μm), which passes pixels with values in a certain range, and sets pixels outside of that range to zero.
- A modified soil-adjusted vegetation index (MSAVI) function (see, e.g., [13]) is applied using data bands 17 (0.519 μm) and 115 (1.296 μm).
- A local texture measure (which we call R5R5, see [4]) is applied to data band 217 (2.325 μm).

A Fisher linear discriminant supervised classifier then finds the optimal linear combination of these bands, given the training data, and a boolean decision threshold is found to maximize detections and minimize false alarms.

From this, we can see that GENIE has in this case been able to identify a small (~1%) relevant subset of the hyperspectral datacube, and has identified a useful mixture of spatial (R5R5) and spectral (SPIKE, MSAVI) processing on those planes. This result, and comparison of it to the results for rice, corn, and other crops, is now of interest for understanding the spatio-spectral signatures of these crops given this imagery, and for the identification of useful subsets of the hyperspectral range of wavelengths for. For example, a multi-spectral imager could be tuned to exploit that feature.

3 Classifying forest with fused imagery

There are a number of previous efforts that combine multispectral and digital elevation model (DEM) data. Bucher and Lehmann [16] use high-resolution multispectral data along with hyperspectral data for land cover classification. The DEM data are used both for orthorectification of the multispectral data sets and for differentiating subclasses of vegetation by height. Zhang, Cassells and van Genderen [17] use fused data from a variety of sources for detection and characterization of underground coal fires in China. Their approach makes use of thermal and multispectral data sources, as well as a DEM, which was used for 3-D visualization of the image data and for deriving depth information about the coal fires. Schistad Solberg, Taxt and Jain [18] explore using Markov Random Fields for multi-source feature extraction, in particular fusing Landsat TM, ERS-1 SAR and GIS for land cover classification. This last reference also gives a good overview of the field.

The data sources used in this work are the U.S. Department of Energy National Nuclear Security Agency's (DOE/NNSA) Multispectral Thermal Imager (MTI) and U.S. Geological Survey (USGS) 1:24k DEM, both of which have been described extensively elsewhere [19-22]. The MTI is among many sensors that produced data of the Los Alamos area during or shortly after the 2000 Cerro Grande/Los Alamos wild fire that devastated ~42,000 acres of forest and scrub land, and destroyed over 200 homes in the town of Los Alamos [10]. The MTI program is now supporting ongoing restoration and analysis work, tracking the effects of mitigation efforts and the slow return of vegetation. The MTI image used here was acquired on January 13, 2002.

The topography of the region of interest is quite complex, ranging from the ~10,000 foot peaks of the eastern wall of the heavily forested Jemez Mountains (a dormant volcano), to the ~7,000 foot narrow mesas and steep canyons on which is located the town and Laboratory of Los Alamos. A number of our target features are naturally linked to altitude, e.g., there is an ecological transition region (ecotone) at approximately 8500 feet separating medium altitude forest dominated by Ponderosa Pine, from high altitude mixed conifer forest (which includes Douglas Fir, White Fir, and Spruce), while the town is located on a group of almost constant elevation mesas. Thus, we expect interesting complementary information in the multispectral imagery and digital elevation data sets.

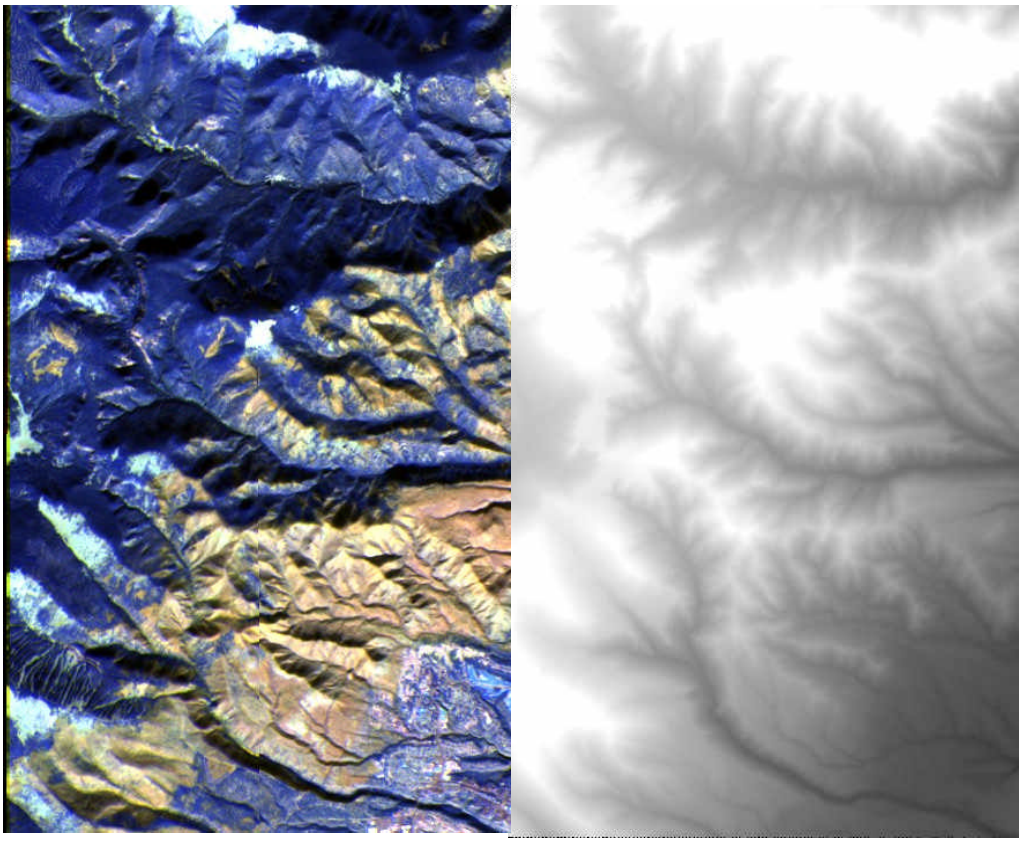

Figure 4. The left panel shows the part of the Jemez Mountains Northwest of the town of Los Alamos. This is a false-color representation of an infrared slice through the Multispectral Thermal Imager (MTI) image cube (using MTI bands O-I-D [19]). Solar angle is such that north-facing slopes appear dark. The right panel shows the matching DEM data, which is co-registered and appended to the multispectral imagery to form our fused datacube. In the DEM, brighter pixels correspond to higher elevations.

Some preprocessing of the data is needed before GENIE can make effective use of them. The MTI instrument team performed calibration and band-to-band registration on their data set. No atmospheric correction is done on the MTI data set. We used the ENVI [23] image processing package to coregister the MTI spectral data and the elevation data contained in the DEM. The coregistration accuracy was less than 3 MTI pixels (i.e., less than 60 meters). This misregistration may be important along the canyon edges, but since the features sought in this work were not expected to depend on small changes in elevation (e.g. along mesa tops and on hillsides), this accuracy was considered sufficient to explore joint MSI/DEM signatures. At this point the coregistered data sets are presented to GENIE for the feature extraction process.

We chose a set of standard land cover classes for which ground truth existed, in the form of an official land cover map [24] for Los Alamos National Laboratory and Los Alamos county. The features we chose to extract were:

- Town/Urban areas
- Wildfire burn scar
- Forest (predominantly Ponderosa Pine, Spruce, Fir, and Aspen)
- Medium elevation Ponderosa pine forest

For each feature described above, a small amount of training data was marked up by hand using a combination of existing land cover maps, first-hand knowledge of the region of interest, and photo-interpretation of the multispectral imagery. A more extensive mark-up of the scene was also prepared to act as out-of-sample test data for each feature. Figure 4 shows the region of interest. Each feature is then extracted, one by one, by GENIE in separate processing runs. Each run required approximately 1 hour of wall-clock time on a standard Linux/Intel workstation. The in-sample (training) and out-of-sample (testing) results for detection rate and false alarm rate for each feature are shown in Table 2.

As an example of the individual results, Figure 5 shows training and the GENIE result for Ponderosa Pine using the fused MSI and DEM data. In each case, the qualitative performance of the algorithm compared to the benchmark manual land cover map is good, and gives confidence that

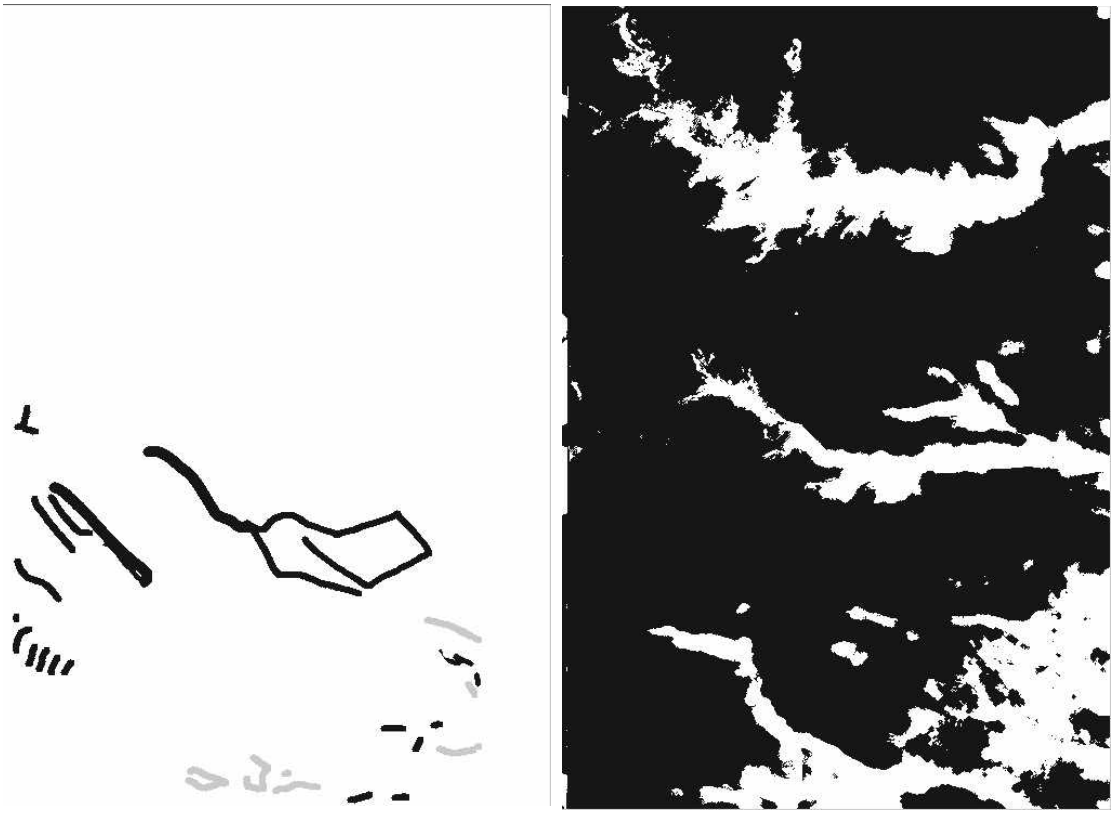

Figure 5. Ponderosa Pine forest feature. Left: Training data provided to GENIE. Black pixels define non-Ponderosa Pine forest training example pixels, and gray pixels define Ponderosa Pine forest example pixels. Right: The GENIE result. Ponderosa Pine forest was detected in pixels marked in white. Compared to the training data, this result achieved a detection rate of 99.9% and a false alarm rate of 0.05%. Outside of the training pixels, performance is qualitatively good, based on comparison to existing, manual land cover maps. The algorithm predominantly detects Ponderosa Pine in the medium elevation valleys and canyons bordering on the Jemez Mountains.

the system is learning valid signatures as opposed to simply over-training on the training data.

The evolved feature extractor with the poorest out-of-sample performance was the town/urban feature extractor. In this case, it appeared that GENIE was only given training data for built-up areas in the town center, and experienced difficulty when tested on urban plus suburban areas. This is understandable, as a large fraction of the suburban component of Los Alamos township is permeated by full-grown trees that fill a substantial aerial fraction when viewed from overhead.

As a test that the system is benefiting from the inclusion of the DEM data, we re-ran the Ponderosa Pine finder problem with the same training data (Fig. 5), but now only presented GENIE with the MTI multispectral imagery. After an equivalent period of training, the performance of the best evolved algorithm (see Table 2) was somewhat less than that of the best algorithm evolved using MSI plus DEM, but the performance outside the training area was noticeably worse, with a substantial decrease in detection rate and a substantial increase in the false alarm rate. In particular, the algorithm trained without access to the DEM data confused Ponderosa Pine forest with high altitude mixed conifer forest throughout the scene.

Table 2. GENIE results for fused MSI and DEM data.

	In-sample Performance		Out-of-sample Performance	
Feature	Detection Rate	False Alarm Rate	Detection Rate	False Alarm Rate
Town	100%	0.16%	78.2%	0.2%
Wildfire Burnscar	100%	0.09%	90.25%	2.2%
Forest	99.4%	0.5%	96.6%	2.3%
Ponderosa Pine	99.9%	0.05%	94.4%	14.7%
Ponderosa Pine without DEM	98.8%	3.70%	83.96%	26.8%

4 Conclusions

We have demonstrated evolution of algorithms on hyperspectral imagery, and on a multi-instrument data set consisting of multispectral (visible to thermal) imagery fused with a digital elevation model (DEM). In seeking to evolve algorithms to extract a range of land cover features, including agricultural crops, town/urban, types of forest, and wildfire burnscars, we find that the system was able to exploit successfully these complex datasets, and produce algorithms that perform well outside the training area. We also demonstrated a case where the same evolutionary system trained to find a particular type of forest, Ponderosa Pine, without the DEM data, had difficulty separating the medium elevation Ponderosa Pine forest from the high elevation mixed conifer forest. We find these results encouraging for future efforts of discovering hyperspectral and multi-instrument signatures of land cover features.

We wish to acknowledge our colleagues on the MTI program from Los Alamos and Sandia National Laboratories and the Savannah River Technology Center, and our colleagues on the ISIS/GENIE team from Los Alamos National Laboratory (LANL). We would like to thank Steve Koch, Leslie Hansen, and Randy Balice of LANL's Ecology Group for access to ground truth data and land cover maps used to prepare our fused training data. This work was supported by the Department of Energy and Department of Defense.

References

[1] S. P. Brumby, J. Theiler, S.J. Perkins, N. R. Harvey, J.J. Szymanski, J.J. Bloch, and M. Mitchell, *Investigation of feature extraction by a genetic algorithm*, Proc. SPIE, Vol. 3812, pp. 24-31,1999.

[2] J. Theiler, N. R. Harvey, S. P. Brumby, J. J. Szymanski, S. Alferink, S. J. Perkins, R. Porter, and J. J. Bloch, *Evolving retrieval algorithms with a genetic programming scheme*, Proc. SPIE, Vol. 3753, pp.416-425, 1999.

[3] S. Perkins, J. Theiler, S. P. Brumby, N. R. Harvey, R. B. Porter, J. J. Szymanski, and J. J. Bloch, *GENIE: A hybrid genetic algorithm for feature classification in multi-spectral images*, Proc. SPIE, Vol. 4120, pp. 52-62, 2000.

[4] N. R. Harvey, J. Theiler, S. P. Brumby, S. Perkins, J. J. Szymanski, J.J. Bloch, R.B. Porter, M. Galassi, and A. C. Young, *Comparison of GENIE and conventional supervised classifiers for multispectral image feature extraction*, IEEE Transactions on Geoscience and Remote Sensing, Vol. 40, pp. 393-404, February 2002.

[5] J. H. Holland, *Adaptation in Natural and Artificial Systems*, University of Michigan, Ann Arbor, 1975.

[6] I. Rechenberg, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*, Fromman-Holzboog, Stuttgart, 1973.

[7] L. Fogel, A. Owens and M. Walsh, *Artificial Intelligence through Simulated Evolution*, Wiley, New York, 1966.

[8] J. R. Koza, *Genetic Programming: On the Programming of Computers by Natural Selection*, MIT, Cambridge, 1992.

[9] N. R. Harvey, S. Perkins, S. P. Brumby, J. Theiler, R. B. Porter, A. C. Young, A. K. Varghese, J .J. Szymanski, and J. J. Bloch, *Finding golf courses: The ultra high tech approach*, Proc. Second European Workshop on Evolutionary Computation in Image Analysis and Signal Processing (EvoIASP2000), Edinburgh, UK, pp. 54-64, 2000.

[10] S. P. Brumby, N. R. Harvey, J. J. Bloch, J. Theiler, S. Perkins, A. C. Young, and J. J. Szymanski, *Evolving forest fire burn severity classification algorithms for multi-spectral imagery*, Proc. SPIE, Vol. 4381, pp. 236-245, 2001.

[11] S. P. Brumby, J. Theiler, J. J. Bloch, N. R. Harvey, S. Perkins, J. J. Szymanski, and A. C. Young, *Evolving land cover classification algorithms for multi-spectral and multi-temporal imagery*, Proc. SPIE, Vol. 4480, pp. 120-129, 2002.

[12] Catherine S. Plesko, Steven P. Brumby and Conway Leovy, *Automatic Feature Extraction for Panchromatic Mars Global Surveyor Mars Orbiter Camera Imagery*, Proc. SPIE, Vol. 4480, pp. 139-146, 2002.

[13] J. A. Richards and X. Jia, *Remote Sensing Digital Image Analysis*, 3rd ed., Chapter 13, Springer, Berlin, 1999.

[14] Peter J. Jarecke, Karen E. Yokoyama, Pamela Barry, *On-orbit solar radiometric calibration of the Hyperion instrument*, Proc. SPIE, Vol. 4480, pp. 225-230, 2002.

[15] R. S. Lunetta and C. D. Elvidge (editors), *Remote sensing change detection*, Ann Arbor, Chelsea, 1998.

[16] T. Bucher and F. Lehmann, *Fusion of HyMap hyperspectral with HRSC-A multispectral and DEM data for geoscientific and environmental applications*, Proc. of IGARSS 2000: IEEE 2000 International Geoscience and Remote Sensing Symposium. Taking the Pulse of the

Planet: The Role of Remote Sensing in Managing the environment, 2000.

[17] X.M. Zhang, C.J.S. Cassells and J.L. van Genderen, *Multi-sensor data fusion for the detection of underground coal fires*, Geologie en Mijnbouw, Vol. 77, pp. 117–127, 1999.

[18] A.H. Schistad Solberg, T. Taxt and A.K. Jain, *A markov random field model for classification of multisource satellite imagery*, IEEE Transactions on Geosience and Remote Sensing, Vol. 34, p. 100, 1996.

[19] W. R. Bell and P. G. Weber, *Multispectral Thermal Imager – Overview*, Proc. SPIE, Vol. 4381, pp. 173-183, 2001.

[20] M. L. Decker and R. Kay, *Multispectral thermal imager satellite hardware status, tasking, and operations*, Proc. SPIE, Vol. 4381, pp. 184-194, 2001.

[21] J. J. Szymanski, W. Atkins, L. Balick, C. C. Borel, W. B. Clodius, W. Christensen, A. B. Davis, J. C. Echohawk, A. Galbraith, K. Hirsch, J. B. Krone, C. Little,P. Mclachlan, A. Morrison, K. Pollock, P. Pope, C. Novak, K. Ramsey, E. Riddle, C. Rohde, D. Roussel-Dupré, B. W. Smith, K. Smith, K. Starkovich, J. Theiler, and P. G. Weber, *MTI Science, Data Products and Ground Data Processing Overview*, Proc SPIE, Vol. 4381, pp. 195-203, 2001.

[22] U.S. Geological Survey Digital Elevation Models: http://edcwww.cr.usgs.gov/glis/hyper/guide/1_dgr_dem

[23] The ENVI software package is described on the web site: http://www.rsinc.com/envi.

[24] S. W. Koch, Ecology Group, Los Alamos National Laboratory, private communication.